Bolt Home Task

Part I Solution

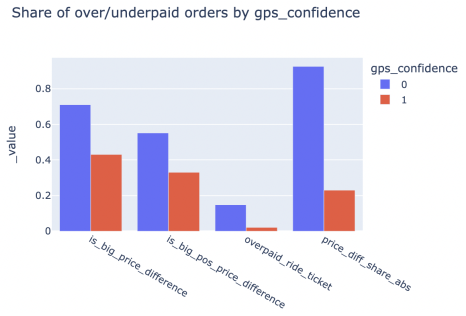
Recommendations:

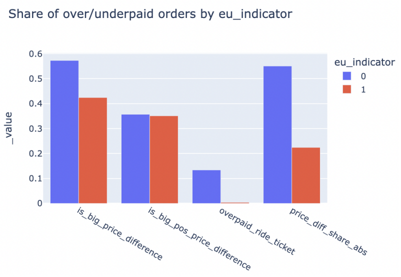
- The top one opportunity to improve precision is considering GPS confidence in model. If it is bad - the price should be higher.
- The second option may be in using indicator of EU/nonEU countries in model. In this case it would be nice to have an additional research about causes of the forecasting errors outside EU. It could be not only GPS connection but also problems with some kind of maps in country or wrong calculation of traffic jams
- The last significant feature is entering destination by driver. In all cases destination changes after it. There also may be pitfalls that require deeper analysis

Insights about features

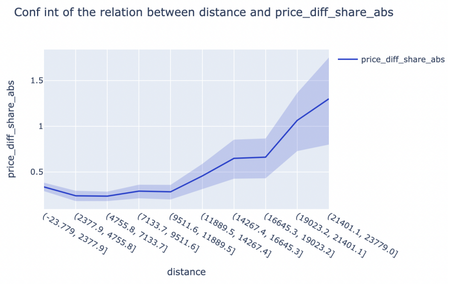
The big price difference is more often in cases when:

- Destination was changed and the driver entered the destination. It is related because in all cases when driver entered the destination it was changed. Share of orders with big price difference (>20%) is 60% vs 45% on average
- GPS connection was bad - share of cases with big diff is 71% vs 45% on average. Bad connection is more frequent in nonEU countries (3% bad conn in EU and 33% in nonEU)
- Country is outside EU - share of cases with big diff is 57% vs 45% on average
- Distance and duration is long in nonEU countries. For distance higher than 15000 share of big price diff grows to ~80-90% and for duration bigger than 2000 it grows to ~66%. This may be the cause of bad gps connection

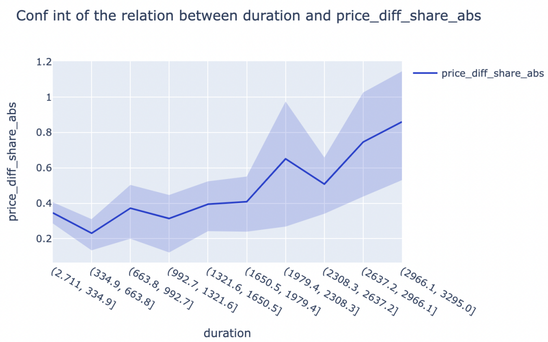




```
EU: 0
Corr: 0.3880879598588232
```



```
EU: 0
Corr: 0.21402385536634802
```

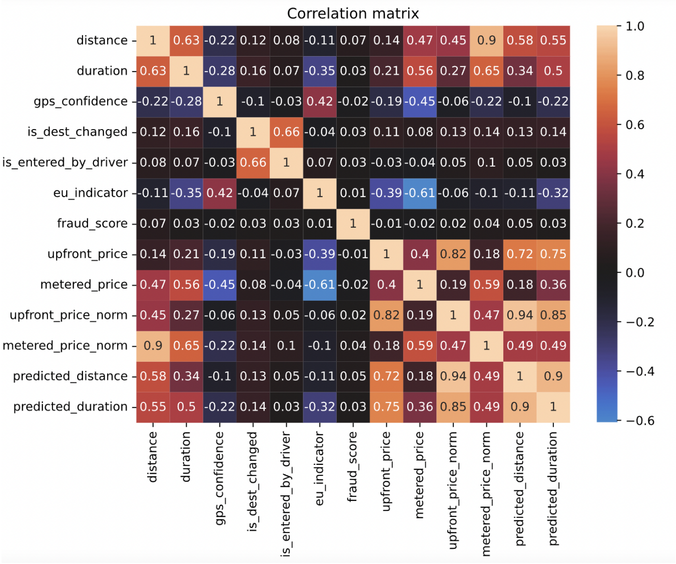


Fraud score doesn't correlate with prediction errors.

Model

We don't know how the previous model was built but can compare the new model based on only destination and duration and model that includes other features.

Correlation between non-target features is less than 0.7 so we can use all of them.



	distance	duration	gps_confidence	is_dest_changed	is_entered_by_driver	eu_indicator	fraud_score	upfront_price	metered_price	upfront_price_norm	metered_price_norm	predicted_distance	predicted_duration
distance	1	0.63	-0.22	0.12	0.08	-0.11	0.07	0.14	0.47	0.45	0.9	0.58	0.55
duration	0.63	1	-0.28	0.16	0.07	-0.35	0.03	0.21	0.56	0.27	0.65	0.34	0.5
gps_confidence	-0.22	-0.28	1	-0.1	-0.03	0.42	-0.02	-0.19	-0.45	-0.06	-0.22	-0.1	-0.22
is_dest_changed	0.12	0.16	-0.1	1	0.66	-0.04	0.03	0.11	0.08	0.13	0.14	0.13	0.14
is_entered_by_driver	0.08	0.07	-0.03	0.66	1	0.07	0.03	-0.03	-0.04	0.05	0.1	0.05	0.03
eu_indicator	-0.11	-0.35	0.42	-0.04	0.07	1	0.01	-0.39	-0.61	-0.06	-0.1	-0.11	-0.32
fraud_score	0.07	0.03	-0.02	0.03	0.03	0.01	1	-0.01	-0.02	0.02	0.04	0.05	0.03
upfront_price	0.14	0.21	-0.19	0.11	-0.03	-0.39	-0.01	1	0.4	0.82	0.18	0.72	0.75
metered_price	0.47	0.56	-0.45	0.08	-0.04	-0.61	-0.02	0.4	1	0.19	0.59	0.18	0.36
upfront_price_norm	0.45	0.27	-0.06	0.13	0.05	-0.06	0.02	0.82	0.19	1	0.47	0.94	0.85
metered_price_norm	0.9	0.65	-0.22	0.14	0.1	-0.1	0.04	0.18	0.59	0.47	1	0.49	0.49
predicted_distance	0.58	0.34	-0.1	0.13	0.05	-0.11	0.05	0.72	0.18	0.94	0.49	1	0.9
predicted_duration	0.55	0.5	-0.22	0.14	0.03	-0.32	0.03	0.75	0.36	0.85	0.49	0.9	1

Feature is_dest_changed is not significant by p-value therefore I exclude it from the model.

Model: 6.8402 + 3.8146*is_entered_by_driver + 1.3109*eu_indicator + 0.001959*predicted_duration + 0.000175*predicted_distance - 4.2905*gps_confidence

The model without new features has R2=0.2568, model with features has R2=0.2886. MSE, RMSE, MAPE and share of big errors (>20%) also got better:

- MSE: 38.2912->36.6551
- RMSE: 6.1880->6.0543
- MAPE: 0.5742->0.5423
- Share all errors: 0.6524->0.6255
- Share overpays: 0.1752->0.1717

This is not a huge increase but it means that we can use features from model for prediction.

The significance of features

Absolute scaled model coefficients:

- predicted_distance 2.022
- predicted_duration 1.599
- gps_confidence 1.278
- eu_indicator 0.550
- is_entered_by_driver 0.487

All those features are significant (by p-value) and coefficients describe the importance of features. So **the top one feature after distance and duration is quality of GPS connection**. It has negative sign so if connection is bad the price gets higher. Other variables has positive signs.